Appendix E – Detailed evidence for ten selected instruments

Each of the ten selected instruments from the systematic review of outcome measures in forensic mental health services is considered in detail below. The Behavioural Status Index is described in greater detail, as a template for considering the remaining tools, which are summarised more briefly.

Behavioural Status Index (BEST)

Background and description of the instrument

The BEST Index is a clinician rated instrument, which is designed to provide an assessment of behaviour. It contains six subscales: insight; communication and social skills; work and recreational activities; self and family care; risk; and empathy. It is intended to be completed by a clinician who has worked with the person being assessed for at least the last six months. It was originally designed for use in a general adult setting to assess change during inpatient treatment and discharge to the community, called the Behavioural Adjustment Index (BAI) (Mahgoub, 1988). In the 1990s it was subsequently adapted for use in forensic psychiatric settings. The original scales were adapted to be more suitable to a forensic context and a risk subscale was added (Robinson et al., 1996). An empathy subscale was also added, based on work by Dutch researchers. The timeframe is the previous 3 months and

the full scale instrument includes 150 items. For each item, raters choose an ordinal response of between 1 and 5. Each response is accompanied by a descriptor for every item, which is listed in the scoring manual.

Overview of the evidence

A total of 13 papers were identified that contained evidence of the psychometric properties of the BEST Index. These papers span a 20 year period from 1999 to 2019. A number of the papers appear to refer to the same study, although this is not made explicit within the papers that they are companion pieces. The results of one study involving 231 participants appear to be published in three papers (Ross et al., 2007, Ross et al., 2008, Ross et al., 2012). The results of another study involving 503 participants appears to be published in seven papers (Woods et al., 2001b, Woods et al., 2001a, Woods et al., 1999, Woods et al., 2003a, Woods et al., 2003b, Woods et al., 2004, Woods et al., 2005) In total there appear to be five separate studies described, involving 1344 participants. Study participants were drawn from England, Scotland, Ireland, Germany, Norway and the Netherlands. There was considerable variability in the form of the tool utilised in these studies. Papers up until 2005 focus on a version with 3 subscales (Insight, Risk, Communication) totalling 70 items. The two studies conducted in The Netherlands examine the performance a 63 item modified version (Chakhssi et al., 2010b, Chakhssi et al., 2010a). This was created from the original 3 subscale, 70 item version following the results of a principal components analysis (PCA), which suggested a four factor solution, containing 63 items. Subsequent analyses were conducted on newly

constituted scales based on original items rearranged in to these four new factors. This means that these results cannot be compared directly with the subscales from the standard version of the BEST. All other studies from 2007 onwards use the full 6 subscale, 150 item version. This version appears to contain the original 3 scales, alongside the 3 new scales. Information about the psychometric properties from the earlier studies should therefore be considered alongside those of the later studies. Information was extracted on tool development, content validity, structural validity, internal consistency, reliability, hypothesis testing and responsiveness. No information was identified on measurement error, measurement invariance or criterion validity.

Instrument development and content validity

The BEST Manual refers to the development of the original four subscales which formed the BAI, which preceded the BEST Index (Reed and Woods, 2002). These evolved as part of a doctoral programme of work (Magoub, 1988). Magoub reports that 'four areas of assessment were chosen to constitute the structure of the BAI'. It is not clear how this selection was made, but it appears to have been on a theoretical basis. Magoub goes on to describe how these subscales were further developed through discussions with 'the researcher, her supervisor, and members of the 1982-83 community psychiatric nursing course in the Department of Health Studies, Sheffield City Polytechnic. The course members included 22 experienced psychiatric nurses'. The 100 item BAI was then tested in a pilot study with 15 Community Psychiatric Nurses (CPN). Subsequently a Recovery Index (RI) was

developed to provide a five point response scale, instead of the previous binary response options.

Robinson et.al. describe the development of the risk subscale of BEST, which was created as an addendum to the original four subscales of communication and social skills, insight, self and family care and work and recreational activity (Robinson et al., 1996). The authors describe the process of developing the risk subscale, which involved a' detailed literature search', as well as 'clinical discussions involving named nurses within four clinical contexts at Rampton Hospital. The go on to describe 'in depth interviews' with named nurses, although the number and nature of the interviews is not described in further detail, other than to comment that the aim was to 'identify, describe and categorise diurnal behaviours associated with or potentially indicative of, risk'. The authors note that 128 behavioural items resulted from this process, which were reduced to 20 after 'categorisation and comparison with the literature'. Again no further information about this process is available.

The BEST manual describes how the empathy subscale was developed using concept mapping techniques with a 'representative sample of multidisciplinary professionals working in the field of forensic mental health care' by the Dutch researchers (Reed and Woods, 2002). No original references are provided which give further details about this development process.

Woods et al. report that 'Face, content and construct validity of the BSI have been established by extensive comparison with normative behavioural models for eye

contact and other social skills'. They go on to report that 'operational versions in the subscales have been further validated by submitting successive drafts to expert clinicians'. This has then been 'validated against the literature and against a survey of clinical opinion among skilled staff working on maximum security wards at Rampton Hospital (Woods et al., 1999). No further description of these processes is presented.

Ross et al. report on the perceived clinical utility of the BEST Index, conducting 102 interviews over 11 clinical sites (Ross et al., 2012). They used a semi-structured interview with 10 prompts and additional probes. The data was analysed based on Grounded Theory. Three themes were identified, namely sharpening staff perception of patient behaviour, enhancing a resource-orientated perspective of treatment and prompting interdisciplinary cooperation and information exchange.

Walker et al. present the results of empirical qualitative enquires with patients and multidisciplinary staff members (Walker et al., 2019). Seventy six patients were interviewed face to face by the same research assistant trained by the study leads. Patient views focused on 'acceptance of the process' with a number of respondents reportedly identifying the importance of being involved in risk assessment. A further 158 staff provided feedback via survey questionnaire, containing questions answered on a Likert scale, as well as free text response boxes. A small number of staff (n=11, 16%) were unfamiliar with the tool or did not understand the meaning of scores. Senior nursing staff are described as being able to see the benefit of the tool. Although this does not focus explicitly on COSMIN's dimensions of

comprehensiveness, relevance and comprehensibility, it provides some limited, indirect information about the content validity of the BEST.

Structural validity

There are two studies which provide information on the structural validity of the BEST Index. In a study involving 503 participants, Woods et al. conducted an exploratory factor analysis on the risk and communication and social skills subscales (Woods et al., 2004) and later on three subscales, adding the insight subscale (Woods et al., 2005). A confirmatory factor analysis was subsequently performed on the Dutch version of the BEST Index (Chakhssi et al., 2010b).

According to the COSMIN Risk of Bias checklist, the quality of the study by Woods et al. was adequate, as an exploratory factor analysis was performed and this included over 7 times the number of items (70) (Woods et al., 2005). This was not a confirmatory factor analysis, so tests were conducted to determine the level of fit, however the authors observe that 'Although the hypothetical model illustrated in the original Venn diagram has not been supported, since there appeared to be no substantial relationship between all three subscales, some interesting and clinically significant themes have emerged'. Chakhssi et al. commented on these results 'In this study, the authors investigated the factor structure of the first three subscales: Risk, Insight, and Social Skills, which resulted in an over factored model consisting of 11 factors, where one factor had no salient loadings and one factor consisted of only one item'. They go on to perform a confirmatory factor analysis and principal

components analysis based on a three factor model proposed in an unpublished theses by Woods (Chakhssi et al., 2010b).

Internal consistency

There is good evidence for internal consistency for all of the original six scales of the BEST index. Additionally there is evidence to support the use of the adapted subscales of the Dutch version, developed by Chakhssi et al., as a result of their factor and principal components analyses (Chakhssi et al., 2010b). There is considerable heterogeneity in the internal consistency scores reported for each of the different subscales.

Woods et al. report internal consistency ratings for three of the subscales, namely risk, insight and communication and social skills (Woods et al., 1999). Internal consistency scores are reported for each of the 70 items in these scales and no overall score is reported for each subscale. The scores are reported as correlations, rather than Cronbach's alpha, and a rho value is given for each item. The authors explain that 'This means that all items are internally consistent with the 'entity' which each subscale purports to measure'. They go on to observe 'Within the risk subscale 19 items correlate very highly ($p<0.001$) and one highly ($p<0.01$). Within both insight subscale and communication and social skills subscale all items correlate very highly ($p<0.001$).' The reported values of rho range from 0.138 to 0.808. This does not meet the quality standards expected by the COSMIN Risk of Bias Checklist, as neither Cronbach's alpha nor Intra-class Correlation are reported.

Ross et al report internal consistencies for all six scales, with alpha values ranging from 0.64 for communication and social skills subscale to 0.73 for the Insight subscale (Ross et al., 2008). Only communication and social skills and social risk subscales have good internal consistency according to the COSMIN Risk of Bias checklist, with alpha values of 0.73 and 0.70 respectively. All the other 4 subscales have alpha score indicative of an inadequate internal consistency. Walker et al. report internal consistencies for all six subscales, with alpha values of 0.90-0.97 (Walker et al., 2019). All of these are considered indications of good internal consistency by the COSMIN Risk of Bias checklist. Despite some heterogeneity, on balance, there is sufficient evidence for good internal consistency across the six subscales.

Reliability

There is good evidence for both inter-rater and test-retest reliability from several studies. Woods et al. report inter-rater and test-retest reliabilities for each item of three of the subscales, namely risk, insight and communication and social skills (Woods et al., 1999). Spearman's rho scores are reported for each of the 70 items in these scales. Test-retest reliability was established in a sample of n=100 and was taken at a two weekly interval, although there is no indication of whether measures were taken to determine if anything had changed in the patients underlying presentation between the two readings. The Spearman's rho was 0.892 for the risk subscale, 0.843 for Insight and 0.878 for communication and social skills. An overall mean agreement for inter-rater reliability is reported as a percentage for each

subscale, as 86% for risk, 82% for insight (n=37) and 78% for communication and social skills (n=35).

Ross et al. report inter-rater reliability scores for all six subscales using 127 double ratings for 30 patients. They report Spearman rank correlation coefficients and intra-class correlations (ICC) for each subscale, with risk r=0.48 and ICC 0.54, insight r=0.57 and ICC 0.58, communication and social skills r=0.48 and ICC 0.47, work and recreational activities r=0.48 and ICC 0.50, self and family care r=0.49 and ICC=0.51, and empathy r=0.52 and ICC=0.51. Walker et al. report ICCs between all subscales 'The average measure intraclass correlation was 0.62, with a 95% confidence interval from 0.38 to 0.76 ($F(99, 495) = 4.66$, $p<0.001$.'

Overall there does not seem to be enough evidence to support adequate reliability. There is limited information provided about the test conditions, stability of patients between testing and the nature of the two reviewers. There is doubt about the appropriateness of some of the statistical methods used, with no kappa scores reported. The ICCs which are reported are often <0.70, indicative of inadequate reliability.

Hypothesis testing

There are a wide range of hypotheses tested through both differences between groups and comparisons with other measures. There are no studies testing the predictive properties of the BEST Index. Woods et al. test several hypotheses for differences by Mental Health Act classification, ward type and gender for the risk and

communication and social skills subscales (Woods et al., 2001b, Woods et al., 2001a). They also examine similar hypothesis for common factors between combinations of two each of the three scales of risk, insight and communication and social skills (Woods et al., 2003b, Woods et al., 2003a, Woods et al., 2004) and for all three scales (Woods et al., 2005). In total they compare 243 subgroups, of which 147 (60.5%) are significant, in line with the hypotheses. Ross et al. reported correlation with four other instruments for each of the six subscales. In total there were 54 correlations calculated, of which 36 (66.7%) were significant (Ross et al., 2008). In Chakhssi et al. the authors report correlations between the original risk, insight, communication and social skills subscales and the total scale and the historical, clinical, risk, total and risk judgement scales of the HCR-20 (Chakhssi et al., 2010b). Out of 20 correlations calculated, 17 (85%) were significant.

Responsiveness

Ross et al. report on the responsiveness of all six subscales of the BEST Index over a period of a minimum of 18 months (Ross et al., 2008). They note that there were significant changes in 3 out of the 6 subscales, namely insight, work and recreation and empathy. Risk, communication and social skills and self and family care did not change over time. They also explored change at an item level and found that only 29 out 150 items (19.3%) showed significant change, although there were items within all six subscales that did show significant changes. The authors note that ‘The reporting period was too short to allow for a reliable measurement of a slowly beginning behavioral change process. The average hospitalization of forensic psychiatric patients is between 5 and 7 years. Because patients tend to change

slowly, most of them are not deemed to be fit for release after shorter periods of treatment.'

Walker et al. report significant improvement after 18 month follow up on 3 out of the 6 subscales, namely risk, insight, and work and recreation. They also note that average empathy scores reduced over time, but not to a significant degree. The authors also note that 'Clinical change (using the marker of 20 per cent improvement in scores) was evident in (n=58, 36%) of the total sample in social risk; (n=60, 39%) in communication; (n=75, 50 per cent) in self-care; (n=77, 48%) in insight; (n=73, 48%) in work and recreation; and (n=69, 46%) in empathy. Clinical deterioration was also evident, but in a smaller proportion of patients: (n=38, 24%) social risk; (n=38, 25%) in communication; (n=35, 23%) in self-care; (n=29, 18%) in insight; (n=35, 25%) in work and recreation; and (n=42, 28%) in empathy. Others patients remained fairly static.'

Summary assessment of the BEST Index

The BEST Index has existed in its current form for over 20 years and its psychometric properties have been extensively studied. Evidence is available about the process of its development and content validity from a number of sources, although this information frequently lacks relevant details, such as numbers of participants and the exact methods of data collection or analysis (Reed and Woods, 2002). Evidence for structural validity provides a mixed picture and is limited to only three of the full six subscales, with confirmatory factor analysis of a three factor model concluding an inadequate level of fit (Chakhssi et al., 2010b). Good internal

consistency is well supported by high Cronbach's alpha scores, but there is some question about how to interpret these without conclusive evidence of good structural validity (Walker et al., 2019, Ross et al., 2008). Evidence of reliability is variable, with many scores suggestive of inadequate agreement (Ross et al., 2008, Woods and Reed, 1999, Walker et al., 2019). Numerous hypotheses are tested, although a sizeable proportion are non-significant. The evidence for responsiveness from two studies is equivocal, with some subscales demonstrating significant change, although these subscales are not consistent between studies (Ross et al., 2008, Walker et al., 2019).

Sexual Violence Risk 20 (SVR-20)

The SVR-20 was developed based on risk factors for sexual violence guided by four principles (Boer, 1997):

1. Empirically related to future sexual violence
2. Practically useful
3. Not discriminatory
4. Parsimonious

A rationale is provided in the manual by the authors for the inclusion of each item. The authors describe a process of identifying risk factors from the literature and then shortening this by combining related risk factors. The risk factors were then grouped in to three major sections, based on the authors' opinion. No specific content validity evidence identified.

The evidence for psychometric properties in a forensic psychiatric context is limited. Five studies were identified that involved forensic psychiatric assessment, although in several cases it was difficult to determine if the population studied was relevant to this review. Only studies that involved participants who were involved in some form of forensic psychiatric treatment, not just a psychiatric assessment, were included. For the two studies located in the United Kingdom the authors indicate that although the assessments were carried out by the local regional secure unit, the subjects were not necessarily mental disordered, but were included because they were participating in an outpatient therapy programme for sexual offenders (Craig et al., 2004, Craig et al., 2006). There are numerous studies that consider the use of the SVR-20 in the context of prison and other settings. These studies are not considered in more detail here.

Overall, there is very limited evidence for its use as an outcome measure in forensic mental health services. There is some evidence for interrater reliability, which appears to be acceptable for individual subscales and total score, but inadequate for the overall risk rating (de Vogel et al., 2004b, de Vries Robbe et al., 2015b). Other studies explore a variety of hypotheses, including differences between subgroups, correlation with other risk assessment tools and predictive abilities relative to a range of violent and sexually violent outcomes (de Vries Robbe et al., 2015b, de Vogel et al., 2004b, Yoon et al., 2011, Craig et al., 2004, Craig et al., 2006). Overall it performs poorly in these categories, with 42/78 (53.8%) hypotheses not supported.

Structured Assessment of Protective Factors for risk of violence (SAPROF)

The SAPROF was first published in Dutch in 2007 (De Vogel et al., 2007). It was subsequently translated in to English and several other languages. It was specifically designed to have a dynamic focus (de Vogel et al., 2009). In the development of the initial 16 factor research version (de Vogel et al., 2004a) the authors aimed to identify protective factors that were:

1. Empirically related to reduced future risk of violence
2. Mainly dynamic, so that they could form targets for treatment
3. A manageable number of items

These items were based on literature reviews of protective factors and contextual factors, as well as the clinical experience of professionals and researchers at the Van der Hoeven Kliniek in the Netherlands. The clinical input consisted of asking a range of mental health professionals who participated in 60 case conferences to suggest factors that may be protective against a relapse in to violent behaviour. Items in the prototype version were reduced by field testing and two that were hard to code were removed. One item was also subdivided in to two items. Item names were also reworded to make them more neutral and less ambiguous. The categorisation of items was also altered from one that mirrored the HCR-20 to the final structure of internal, motivational and external subscales (de Vogel et al., 2012).

In total 12 studies were identified that contained information about the psychometric properties of SAPROF in a forensic psychiatric context. The evidence for its

psychometric properties as an outcome measure are strong in a number of areas, but absent in others areas. There are is no evidence to support good structural validity and the three subscales have not been empirically validated. There are two studies of adequate quality that indicate good internal consistency for the overall scale (Abidin et al., 2013, Kashiwagi et al., 2018). There is no evidence for the internal consistency of the subscales, although this would be appropriate given the lack of evidence for adequate structural validity. There is evidence from 8 studies for inter-rater reliability. Despite some inconsistency, overall there is strong evidence for adequate reliability, especially of the total score. No studies were identified that examined test-retest reliability. Hypothesis testing was conducted in all of the included studies, including predictive validity (mainly for violence), difference between subgroups and correlation with a number of other measures, particularly the HCR-20. In total 199 out of 321 (62%) results were in line with the hypotheses. Responsiveness was considered by two studies which compared pre and post treatment scores over a variable period of follow up (de Vries Robbe et al., 2011, de Vries Robbe et al., 2015a). In total 14 out of 14 (100%) scores showed significant change after treatment.

Violence Risk Scale (VRS)

The Violence Risk Scale (VRS) was designed to be a generic risk assessment for ‘forensic clients’, in particular those that were ‘being considered for release from institutions to the community after a period of treatment’ (Wong and Gordon, 2000). The VRS is deliberately designed to measure change in an individual’s risk of

violence over time, and therefore to provide an assessment of the effectiveness of treatment in reducing this risk. Those dynamic risk items that have high scores are potential targets for intervention. The 20 dynamic risk factors included in the tool are derived from 'risk assessment and treatment literature and are empirically or theoretically linked to violence'. The VRS is based on the Transtheoretical Model of Change proposed by Prochaska and DiClemente (Prochaska and DiClemente, 1982). This aims to identify the level of readiness for change in the individual undergoing the assessment. The score on each dynamic item corresponds to a level on the model of change i.e. pre-contemplation/contemplation, preparation, action and maintenance. The tool is also based on the psychology of criminal conduct and the principles of effective correctional treatment (Wong and Gordon, 2006). The use of the VRS consists of a part A which is administered pre-treatment and consists of both static and dynamic measures and part B, which measures response to treatment and consists solely of the dynamic factors (Dolan et al., 2008).

There were 13 papers included in the review, involving at least 1852 participants. There appeared to be at least some overlap in the samples of two of the studies, so these numbers are not double counted, although it is not clear the degree of overlap (Lewis et al., 2013, Olver et al., 2013). Three studies took place in a sample of offenders being treated in a psychiatric environment in Alberta, Canada (Coupland and Olver, 2018, Olver et al., 2013, Lewis et al., 2013). Although it is not clear if these were forensic psychiatric patients, these studies were included as the participants had a history of violence towards others and were receiving psychiatric treatment. One study in general psychiatric patients did not appear to include any forensic patients and so it was excluded (Doyle et al., 2012)

Evidence for structural validity is limited to exploratory factor analyses performed by the tool's authors (Wong and Gordon, 2006). They conclude that 'Results of the exploratory factor analyses provide some evidence for the separation of the static and dynamic variables. The results also suggest that the static variables lack unidimensionality; three of the six static variables were loaded on Factors 1 and 3, and this may account for the low alpha (internal consistency) of the static variables'. There is no confirmatory factor analysis. The same study is the only evidence of internal consistency, which appears to be good for the total and dynamic scores. Evidence of good interrater reliability is available from multiple studies. It is also reported from an unpublished source by Wong and Gordon (Wong and Gordon, 2006). It is reported that '60 cases from the same total sample based on file and interview information. Pearson correlation of the two ratings was 0.87' (Gordon, 1998). Four studies consider the measurement error of the instrument, but only one uses a clinical, rather than just statistical, methodology (Howden et al., 2018). None meet COSMIN criteria for adequate evidence (Mokkink et al., 2018). Numerous hypotheses are tested in 11 different studies, including difference between subgroups, prediction of outcomes and correlation with other measures. In total 223/334 (67 %) of hypotheses were supported. There is evidence regarding responsiveness from 8 studies which give a mixed picture. Some suggest statistically significant change (Coupland and Olver, 2018, Hogan and Olver, 2016, Wilson et al., 2014), but those considering change indices suggest either limited or no improvement (Howden et al., 2018, Draycott et al., 2012). One study did suggest significant change that was also reliable for a majority of participants (Horgan et al., 2019).

Level of Service: Case Management Inventory (LS/CMI)

The Level of Service: Case Management Inventory is the most recent iteration of a series of tools in the Level of Service Inventory family. These tools are based on the Risk-Needs-Responsivity theory of criminal rehabilitation (Andrews and Bonta, 2010). This has three underlying principles which guide the approach:

(1) target resources to those offenders posing a higher risk

(2) focus interventions on criminogenic needs

(3) consider individuals' abilities when tailoring interventions

The LS/CMI is designed to be a 'comprehensive measure of risk and need factors, as well as a fully functional case management tool' (Andrews et al., 2004). The LS/CMI is described by its authors as a fourth generation risk assessment because it integrates these different elements with the aim of addressing risks to reduce recidivism. It is an evolution of earlier tools, such as the Level of Service Inventory-Revised (LSI-R) and incorporates many elements of its predecessors. The LS/CMI purports to combine and refine the 54 items of the LSI-R in to the 43 items of the LS/CMI section 1. The LS/CMI items are very similar to those in the LSI-R items. The subscales are reduced to 8 from 10, with those covering accommodations, financial and emotional/personal items removed and a new subscale on antisocial pattern added (Jung et al., 2012). The LS/CMI is intended for use in a wide range of criminal justice settings, including prison and probation. It is not explicitly developed for use in healthcare settings, such as forensic mental health hospitals or community services. Mental health is only considered in section 4 'other client issues', which does not contribute to the overall score, but can be used for administrative override purposes

and to inform case management. The tool can be completed by a variety of professionals, including health care workers, such as physicians and psychologists, and those involved in criminal justice and social work, such as probation officers and youth workers. It is available in both paper and software formats. The level of service tools are based on static and dynamic risk factors for recidivism. The LS/CMI revision also relies on 'user consultation'.

There is very limited published evidence about the use of the LS/CMI in a forensic mental health setting. In total only three studies were identified and all of these only considered the general scales in section 1 (see appendix D). Two of the studies were set in an outpatient psychiatric facility involving participants who were offenders referred for psychiatric assessment (Jung et al., 2012, Jung et al., 2013). These studies also included some data from the LSI-R, but omitted those items which were not in the LS/CMI. This means that conclusions can be drawn from this data about the performance of the LS/CMI subscales that also form part of the LSI-R. Evidence was only available for internal consistency and hypothesis testing. Internal consistency was only considered in one study and only 6 out of the 8 subscales of the general risk/needs scale were calculated101 (Jung et al., 2012). Internal consistency varied considerably between subscales from 0.80 for criminal history to just 0.07 for antisocial orientation. Hypothesis testing was conducted in all three included studies, but was limited to correlations with other measures, including the CANFOR, HoNOS Secure and HCR-20. Out of 193 hypotheses tested, 63 (35%) were supported with significant correlations. Additional evidence derived from the more established LSI-R and from other populations, such as those in the criminal

justice system, would provide additional tangential clues about the performance of the LS/CMI in this population.

Health of the Nation Outcomes Scales – Secure (HoNOS-Secure)

The Health of the Nation Outcome Scales –Secure is part of the HoNOS family of outcome measurement instruments. The different versions are designed to be used in different ages or clinical populations. HoNOS are clinician rated instruments that originated in the United Kingdom, but are now used in many countries around the world. The original HoNOS scale was developed in the 1990s by the Royal College of Psychiatrists' Research Unit. It was created in response to a target set by the UK's Department of Health to improve the health and social function of people with mental illness (Department of Health, 1992). The developers aimed to create a tool that was short enough to be used by 'keyworkers' that covered both clinical problems and social functioning. They also wanted it to be able to reliably detect change, as well as demonstrate a relationship to scales that were more established at that time (Wing et al., 1998). The measure was developed through an iterative process that involved systematically reviewing the literature and widespread consultation. The literature review revealed a large number of scales, but none that were deemed sufficiently broad in scope and brief to complete. The authors therefore initiated the process of developing the tool, which was subsequently tested and modified during a four-stage process. The first version (HoNOS-1) was created in consultation with 'experts' although no further evidence is published about the number or type of these consultants. This 20 item pilot measure was tested in a cohort of 152 patients in 9 sites to explore its acceptability, structure and sensitivity to change. Following

feedback from users, it was shortened to 12 items. The resultant second version (HoNOS-2) was tested in a further 100 participants in 7 sites. This resulted in a HoNOS-3, which was then tested in 2706 participants in 25 sites for acceptability, structure, sensitivity, reliability and clinical profiles. The results were used to modify the tool again to produce the final version of HoNOS, which was tested again in 641 participants in 6 sites.

The first version of the HoNOS for forensic services was developed in 2002 within a group of independent hospitals, now called St. Andrew's Healthcare. This was initially known at the Mentally Disordered Offender (MDO) scale. It constituted an additional 7 items, which formed a security scale, rated alongside the original 12 items from HoNOS (Royal College of Psychiatrists, 2021). This was reported to correlate highly with the original HoNOS is an unpublished study in 5 secure units (Dickens et al., 2010). The second version was subsequently developed using 'qualitative consultation and case vignettes in order to establish face and consensual validity' and was published in 2004 (Dickens et al., 2007).

There is extensive evidence for the use of HoNOS Secure in a forensic psychiatric population, with a total of 21 studies were identified involving 2440 participants (see appendix D). No studies were identified examining the structural validity of the HoNOS Secure. Despite this there were two studies that examined the internal consistency of the two HoNOS subscales (Dickens et al., 2007, Dickens and O'Shea, 2017). The values of Cronbach's α suggest good internal validity, however given the lack of evidence for structural validity, these results were assumed to be of

indeterminate quality. Only one study examined the interrater reliability of the HoNOS Secure (Dickens et al., 2007). Despite being of adequate methodological quality, the study only examined individual items and not the reliability for the total or subscale scores. The values of the ICC were highly variable, with some items demonstrating good reliability and other poor reliability. Test-retest reliability is referred to as being 0.65 (Long et al., 2011), which references (Dickens et al., 2007), although there is no mention of this in the original paper. Only one study considers measurement error and this is calculated only by statistical methods (Longdon et al., 2017). Hypothesis testing is undertaken in 14 studies and includes prediction of violence, difference between subgroups such as gender, security level and legal status and correlation with other measures, including risk assessments, neuropsychological measures and measures of social functioning. Overall 103 out of 268 hypotheses (38%) tested are substantiated. Prediction is only tested in one of these studies, but the AUCs suggest excellent predictive abilities (Finch et al., 2017). Responsiveness is examined in 11 studies. The resultant picture is mixed, with a total of 58 out of 120 pairings (48%) showing significant change. This is supported by the mixture of improvement and deterioration observed in the two studies considering clinically important change (Dickens and O'Shea, 2017, Longdon et al., 2017).

Dangerousness Understanding, Recovery and Urgency Manual (DUNDRUM)

The Dangerousness Understanding, Recovery and Urgency Manual (DUNDRUM) is a linked toolkit of rating scales designed to be used a different parts of the forensic mental health pathway (Kennedy et al., 2010). Each scale is designed as a structured professional judgement instrument to aid clinical decision making, in particular guiding the placement of patients at the appropriate level of security. The authors caution that the DUNDRUM toolkit is not meant to replace risk assessment and recommend that it is used alongside the HCR-20. The original toolkit consisted of four clinician rated scales, named DUNDRUM 1 Triage Security, DUNDRUM 2 Triage Urgency, DUNDRUM 3 Programme Completion and DUNDRUM 4 Recovery Items. DUNDRUM 1 and 2 are intended to be used prior to admission to determine the level of security at which patients enter forensic mental health services. As the DUNDRUM 1 and 2 are not measures of the outcome of care they are not assessed in this review. DUNDRUM 3 and 4 are designed to guide moves along the recovery pathway, to make decisions about readiness for a move to less secure inpatient setting or discharge to the community. Patient reported versions of DUNDRUM 3 and 4 were subsequently developed, which mirror the items in the original clinician version (Davoren et al., 2015). The response options are designed to correspond to levels of security, ranging from the community to high secure inpatient services. The authors explain that, as it is a structured professional judgement tool, these are just a guide to decision making and the ultimate choice remains within the clinician's remit.

The DUNDRUM quartet was developed by a team at the Central Mental Hospital, which is the only forensic psychiatric inpatient unit in the Republic of Ireland. The items in the scale are reportedly drafted based on a number of previously developed assessment criteria, decision algorithms and structured professional judgements along with 'our own experience and research'. The DUNDRUM 3 and 4 are also based on existing scales such as the HCR-20, CANFOR and HoNOS and a number of theoretical models, including Maslow's hierarchy of need, engagement, recovery and the trans-theoretical stages of change (Richter et al., 2018). The authors explain that for the DUNDRUM manual 'many colleagues have contributed to this text through comments, criticisms and feedback', although this process is not described in further detail. The content of the patient reported scales of the DUNDRUM 3 and 4 are described as being developed in consultation with one service user, in order to 'allow ease of interpretation, while ensuring fidelity to the clinician rated items' (Davoren et al., 2015). The process of consultation is not described in detail. The self-rated scales have been published in a later edition of the DUNDRUM manual (Kennedy et al., 2013).

There is a considerable amount of evidence for the psychometric properties of the DUNDRUM 3 and 4. In total 8 studies were identified, involving up to 967 participants (see appendix D). The majority of these studies are conducted by the team that developed the quartet, with 6 out of the 8 studies identified taking place in the Central Mental Hospital in Dundrum, Republic of Ireland. It appears likely that there is significant overlap in the study populations for several of the earlier studies in the Central Mental Hospital, with sampling frames either directly overlapping or temporally close, often involving the majority of patients in the hospital at that time.

Structural validity is examined in one study through a principal components factor analysis (O'Dwyer et al., 2011). The results support the unidimensionality of the DUNDRUM 3 and 4 scales, however the sample of 95 is just below the expected minimum of 100 stipulated by COMSIN. No confirmatory factor analysis was performed. Internal consistency is examined in four studies and the evidence all demonstrates high values of Cronbach's α for both the clinician and patient rated versions of DUNDRUM 3 and 4. Evidence for the reliability of the DUNDRUM 3 and 4 is limited, with only one study reporting values for inter-rater reliability (O'Dwyer et al., 2011). There is no evidence of test-retest reliability. For the DUNDRUM 3 only 1 study of measurement error was identified, which calculated a reliable change index (RCI) using statistical methods (Richter et al., 2018). This study compared it to a measure of clinically meaningful change based on the theoretical basis of the instrument, linked to levels of security, rather than empirical qualitative methods. There was extensive testing of hypotheses, including the difference between subgroups such as those with leave and those without leave, correlation with a range of other instruments, such as the HCR-20, CANFOR and SAPROF and prediction of violence, self-harm and moves within the patient pathway. In total 145 out of 192 (76%) of hypotheses identified were supported. There is one study of responsiveness, which considers the DUNDRUM 3 scale only (Richter et al., 2018). Overall the total sample shows significant change, however there is a mixed picture when this is broken down into two subgroups of longer and shorter stay patients. The proportion of patients showing change greater that the RCI and clinically significant change are also calculated. No evidence for the responsiveness of the DUNDRUM 4 or the two patient self-reported scale was identified.

Camberwell Assessment of Need – Forensic Version (CANFOR)

The Camberwell Assessment of Need – Forensic Version (CANFOR) is part of a family of needs assessment instruments designed to be used with different populations of mental health service users. The original version, the Camberwell Assessment of Need (CAN) was developed in the 1990s by a team at the Institute of Psychiatry at King's College London (Phelan et al., 1995). It was designed specifically to assist local authorities to fulfil their obligations under the National Health Service and Community Care Act 1990. The CAN was designed to be quick and easy to use and to be relevant to anyone with severe mental illness. An initial draft was created by the authors, which was then sent to 50 mental health professionals and 59 service users for feedback. Two additional items were added as a result of professional feedback, but none as a result of the feedback from service users. Service users rated all items at least moderately important.

The CAN was used as a template to develop the CANFOR (Thomas et al., 2008). The authors identified that although many domains in the CAN were relevant to forensic mental health service users, many were not covered in sufficient depth. The domains were therefore reworded and additional domains were added. This process was carried out by a team of five professionals from different disciplines. This draft was then piloted with 20 service users and 17 staff members and revisions were made based on the feedback obtained. Content validity was then investigated by interviewing 60 services users, who were asked to rate the relevance of each item on a four point scale. Two additional items were suggested, but these referred to

intervention, not needs, so were not further considered. Fifty professionals were also surveyed to ascertain their views on the need for the CANFOR, the relevance of the ratings, the length of the scale and its comprehensiveness. Comprehensibility was investigated through the application of the Flesch ease of reading score, on which the CANFOR scored 59, indicating that it was at the appropriate level for most readers.

The evidence for the psychometric properties of CANFOR was limited to reliability and hypothesis testing. A total of 10 studies were identified, involving up to 794 patient participants. Reliability was investigated in four studies, which all examined both inter-rater and test retest reliabilities. Overall the methods appear appropriate and the values of ICCs and Cohen's κ indicate adequate reliability, at least at the level of aggregate scale scores. Hypothesis testing covers difference between subgroups, such as patients in decreasingly secure settings and correlation with other instruments. There are no studies that look at the predictive ability of CANFOR. In total 59 out of 96 (61%) of hypotheses are supported.

Short Term Assessment of Risk and Treatability (START)

The Short Term Assessment of Risk and Treatability (START) was developed by a team based in Canada, led by Professor Christopher Webster, who was also closely involved in the creation of the HCR-20. It was based on an earlier tool called the Short-Term Assessment of Risk (STAR) (Webster et al., 2006). The development of the STAR was led by a four person team of professionals from different disciplines,

including mental health nurses, a psychologist and a researcher. The process is described as an iterative one, initially driven by the 'self-reflective process' of one of the mental health nurses involved. A list of '50 or so variables' derived in this way was refined by the research team, in consultation with other professionals and in response to weekly clinical meetings in the minimum security service of St. Joseph's Healthcare in Hamilton, Ontario in Canada. An unstructured process of reviewing the literature pertinent to each item of the evolving tool was undertaken in parallel. The STAR was designed to give equal weight to risk and protective factors. After 6 months, the STAR was published in a draft form (Middleton et al., 2002). The core research group was subsequently expanded to include a second site, also located in Canada, and led to the incorporation of the concept of treatability. The first comprehensive manual for the START was published, which also included an abbreviated guide and a summary sheet (Webster et al., 2004). A revised version of the manual was subsequently published, which claimed not to be designed to replace the original (Webster et al., 2009). The content of three items was altered and some of the 'setting quotations' were also changed.

There is a considerable amount of evidence from a range of different settings to support the content validity of START. A study of 12 staff members in a medium secure unit in the UK found that several participants felt START was useful to organise information about a patient, but raised concerns about uncertainty of the timeframe and the subjectivity of assessments (Doyle et al., 2008). A systematic review of the START identified seven studies that considered feasibility and utility (O'Shea and Dickens, 2014). The authors concluded that overall users were positive about the START, with between 62 and 92.5% endorsing the START's utility.

A total of 28 studies were identified, which contained information on the START's other psychometric properties (see appendix D). These studies contained up to 4740 participants, although it was unclear in one study if the number referred to participants or assessments (Nicholls et al., 2011). One study used an earlier version of the START, which contained the same items as later versions, but was rated on a continuous six point scale from a very high risk to very high strength, rather than two separate scales for strengths and vulnerabilities (Nicholls et al., 2006). There were no studies that explored the structural validity of the START. There were 5 high quality studies that considered internal consistency, with Cronbach's α score consistently >0.70, indicating a high consistency. The lack of evidence for structural validity does however cast doubt on whether the START scales are unidimensional. A total of 10 studies were identified that contained information about reliability. Most focused on inter-rater reliability, but 1 did consider test-retest reliability (Whittington et al., 2014). There was considerable variability in both the quality and results of the studies, with ICC values ranging from 0.30-0.95. All 28 studies tested hypotheses, including difference between subgroups, such as level of security and sex, correlation with other measures, such as HoNOS Secure and HCR-20, and prediction of a range of outcomes, including violence, self-harm and victimisation. The results were highly variable between studies. A total of 1256 hypotheses were tested, with 601 (48%) being supported. Only 2 studies considered responsiveness, with one of poor quality (Nonstad et al., 2010). The second study was of good quality and concluded that the START showed adequate responsiveness (Hogan and Olver, 2016).

Historical, Clinical, Risk – 20 (HCR-20) Version 3

The Historical, Clinical, Risk – 20 (HCR-20) Version 3 is the latest incarnation of the well-established HCR tools, following on from version 1 (1995) and version 2 (1997) (Douglas et al., 2013). The first version was developed by a team based in Canada, based on evidence derived from systematic reviews of risk factors for violence. Its development involved consultation with front line practitioners. Feedback from empirical studies identified that the description of administrative procedures, especially the definition of some risk factors was unclear. The second version of the HCR-20 retained the same 20 items, with updated descriptions. This was extensively studied by research teams in many countries, with numerous systematic reviews supporting its psychometric qualities, including interrater reliability, concurrent validity and predictive validity (Campbell et al., 2009, Singh et al., 2011, Douglas and Reeves, 2010).

The authors developed a draft of the third version of the HCR-20 in 2008, in order to incorporate the latest research since the previous versions and improve its clinical utility. The manual alludes to discussions with 'several groups of colleagues to provide us with their opinions', but does not provide more details. The authors also commissioned a review of the violence literature in 2007. The first draft was presented at a conference to 'approximately 20 psychiatrists, psychologists, nurses, and researchers' resulting in a number of modifications to clarify areas that were unclear. Beta testing of the draft took place in three sites, all in different countries, where small groups of clinicians ranging from 5-15 provided unstructured feedback.

Draft 2 incorporated a number of significant changes, such as removing the four level response options in favour of the familiar three levels. Draft 3 was extensively piloted in six different countries, with studies examining interrater reliability, concurrent validity and predictive validity. The final published version of version 3 contains the same items as drafts 2 and 3. The manual makes reference to content validity, by stating that this 'rests primarily on the adequacy of the literature reviews on which it is based', rather than on the perceived importance of items by users of the HCR-20. There is some limited evidence that users find version 3 an improvement over previous versions, but no structured qualitative evidence for the content validity of HCR 20 as an outcome measure was identified (Bjorkly et al., 2014, de Vogel et al., 2014).

The evidence for the other psychometric properties of the HCR-20 was limited to the final published iteration of version 3 and does not include the evidence presented on earlier drafts in the manual (Douglas et al., 2013). A total of 18 studies contained relevant evidence, including up to 2340 participants (see appendix D). No studies were identified that consider the structural validity of the HCR-20. Evidence for internal consistency was derived from 4 studies, at least 3 of which were of good quality. The values of Cronbach α were variable, ranging from 0.33 to 0.87. Reliability was examined in 10 studies, at least 8 of which were of good quality. Only interrater reliability was considered. Despite some variability, overall the ICC scores were mostly high. Hypothesis testing was carried out in all but 1 of the studies, including difference between subgroups, such as sex and violence, correlation with other measures, such as the START and prediction of a range of outcomes, including violence. A total of 682 hypotheses were identified, of which 341 (50%)

were supported. Responsiveness was reported in 3 studies, all of good quality. Clinical and risk scales showed evidence of significant change in 2 out of the 3 studies.

References

Abidin, Z., Davoren, M., Naughton, L., Gibbons, O., Nulty, A., & Kennedy, H. 2013. Susceptibility (risk and protective) factors for in-patient violence and self-harm: Prospective study of structured professional judgement instruments START and SAPROF, DUNDRUM-3 and DUNDRUM-4 in forensic mental health services. BMC Psychiatry, 13, 197.

Andrews, D., Bonta, J. & Wormith, S. 2004. The Level of Service/Case Management Inventory (LS/CMI) technical brochure. Toronto, Canada: Multi-Health Systems.

Andrews, D. & Bonta, J. 2010. Rehabilitating criminal justice policy and practice. Psychology, Public Policy, and Law, 16, 39-55.

Bjorkly, S., Eidhammer, G. & Selmer, L. 2014. Concurrent validity and clinical utility of the HCR-20 v3 compared with the HCR-20 in forensic mental health nursing: Similar tools but improved method. Journal of Forensic Nursing, 10, 234-242.

Boer, D. 1997. Manual for the Sexual Violence Risk-20: Professional guidelines for assessing risk of sexual violence. Vancouver, Canada: British Columbia Institute Against Family Violence.

Campbell, M., French, S. & Gendreau, P. 2009. The prediction of violence in adult offenders a meta-analytic comparison of instruments and methods of assessment. Criminal Justice and Behavior, 36, 567-590.

Chakhssi, F., de Ruiter, C. & Bernstein, D. 2010a. Change during forensic treatment in psychopathic versus nonpsychopathic offenders. Journal of Forensic Psychiatry & Psychology, 21, 660-682.

Chakhssi, F., de Ruiter, C. & Bernstein, D. 2010b. Reliability and validity of the dutch version of the Behavioural Status Index: A nurse-rated forensic assessment tool. Assessment, 17, 58-69.

Coid, J., Kallis, C., Doyle, M., Shaw, J. & Ullrich, S. 2015. Identifying causal risk factors for violence among discharged patients. PLoS ONE, 10, e0142493.

Coupland, R. & Olver, M. 2018. Assessing dynamic violence risk in a high-risk treated sample of violent offenders. Assessment, 27, 1886-1900.

Craig, L., Beech, A. & Browne, K. 2006. Cross-validation of the Risk Matrix 2000 sexual and violent scales. Journal of Interpersonal Violence, 21, 612-633.

Craig, L., Browne, K. & Stringer, I. 2004. Comparing sex offender risk assessment measures on a uk sample. International Journal of Offender Therapy and Comparative Criminology, 48, 7-27.

Davoren, M., Hennessy, S., Conway, C., Marrinan, S., Gill, P. & Kennedy, H. 2015. Recovery and concordance in a secure forensic psychiatry hospital - the self rated DUNDRUM-3 programme completion and dundrum-4 recovery scales. BMC Psychiatry, 13, 185.

de Vogel, V., de Ruiter, C. & Bouman, Y. 2004a. Scoringshandleiding SAPROF. Structured assessment of protective factors for violence risk. Research versie. [SAPROF coding manual. Structured assessment of protective factors for violence risk. Research version.]. Utrecht/Nijmegen, The Netherlands: Van der Hoeven Kliniek/Trimbos Instituut/Pompestichting.

de Vogel, V., de Ruiter, C., Bouman, Y. & de Vries Robbe, M. 2007. SAPROF: Richtlijnen voor het beoordelen van beschermende factoren voor gewelddadig gedrag. Versie 1. [SAPROF. Guidelines for the assessment of protective factors for violence risk. Version 1]. Utrecht, The Netherlands: Forum Educatief.

de Vogel, V., de Ruiter, C., Bouman, Y. & de Vries Robbe, M. 2009. SAPROF: Guidelines for the assessment of protective factors for violence risk [English version of the Dutch original]. Utrecht, The Netherlands., Forum Educatief.

de Vogel, V., de Ruiter, C., Bouman, Y. & de Vries Robbe, M. 2012. Structured assessment of protective factors for risk of violence (SAPROF): Guidelines for the assessment of protective factors for violence risk. Utrecht, The Netherlands: Forum Educatief.

de Vogel, V., de Ruiter, C., van Beek, D. & Mead, G. 2004b. Predictive validity of the SVR-20 and Static-99 in a Dutch sample of treated sex offenders. Law and Human Behavior, 28, 235-251.

de Vogel, V., van den Broek, E. & de Vries Robbe, M. 2014. The use of the HCR-20 V3 in Dutch forensic psychiatric practice. International Journal of Forensic Mental Health, 13, 109-121.

de Vries Robbe, M., de Vogel, V. & de Spa, E. 2011. Protective factors for violence risk in forensic psychiatric patients: A retrospective validation study of the SAPROF. International Journal of Forensic Mental Health, 10, 178-186.

de Vries Robbe, M., de Vogel, V., Douglas, K. & Nijman, H. L. 2015a. Changes in dynamic risk and protective factors for violence during inpatient forensic psychiatric treatment: Predicting reductions in postdischarge community recidivism. Law and Human Behavior, 39, 53-61.

de Vries Robbe, M., de Vogel, V., Koster, K. & Bogaerts, S. 2015b. Assessing protective factors for sexually violent offending with the saprof. Sexual Abuse, 27, 51-70.

Department of Health 1992. Health of the Nation: White Paper. London, UK: Department of Health.

Dickens, G., Sugarman, P., Picchioni, M. & Long, C. 2010. HoNOS-Secure: Tracking risk and recovery for men in secure care. British Journal of Forensic Practice, 12, 36-46.

Dickens, G., Sugarman, P. & Walker, L. 2007. HoNOS-Secure: A reliable outcome measure for users of secure and forensic mental health services. Journal of Forensic Psychiatry and Psychology, 18, 507-514.

Dickens, G. & O'Shea, L. 2017. Reliable and clinically significant change in outcomes for forensic mental health inpatients: Use of the HoNOS-Secure. International Journal of Forensic Mental Health, 16, 161-171.

Dolan, M., Fullam, R., Logan, C. & Davies, G. 2008. The Violence Risk Scale second edition (VRS-2) as a predictor of institutional violence in a British forensic inpatient sample. Psychiatry Research, 158, 55-65.

Douglas, K., Hart, S., Webster, C. & Belfrage, H. 2013. HCR-20 V3 assessing risk for violence: User guide. Burnaby, Canada: Mental Health, Law, and Policy Institute, Simon Franser University.

Douglas, K. & Reeves, K. 2010. Historical-Clinical-Risk management-20 (HCR-20) violence risk assessment scheme: Rationale, application, and empirical overview. IN: Otto, R. & Douglas, K. [eds.] Handbook of Violence Risk Assessment. New York City, USA: Routledge/Taylor & Francis Group, pp.147-185.

Doyle, M., Carter, S., Shaw, J. & Dolan, M. 2012. Predicting community violence from patients discharged from acute mental health units in england. Social Psychiatry and Psychiatric Epidemiology, 47, 627-637.

Doyle, M., Lewis, G. & Brisbane, M. 2008. Implementing the Short-Term Assessment of Risk and Treatability (START) in a forensic mental health service. Psychiatric Bulletin, 32, 406-408.

Draycott, S., Kirkpatrick, T. & Askari, R. 2012. An idiographic examination of patient progress in the treatment of dangerous and severe personality disorder: A reliable change index approach. Journal of Forensic Psychiatry and Psychology, 23, 108-124.

Finch, B., Gilligan, D., Halpin, S. & Valentine, M. 2017. The short- to medium-term predictive validity of static and dynamic risk-of-violence measures in medium- to low-secure forensic and civil inpatients. Psychiatry, Psychology and Law, 24, 410-427.

Gordon, A. 1998. The interrater reliability, internal consistency and validity of the Violence Risk Scale—experimental version 1. Saskatoon, Canada: University of Saskatchewan.

Hogan, N. & Olver, M. 2016. Assessing risk for aggression in forensic psychiatric inpatients: An examination of five measures. Law and Human Behavior, 40, 233-243.

Horgan, H., Charteris, C. & Ambrose, D. 2019. The violence reduction programme: An exploration of posttreatment risk reduction in a specialist medium-secure unit. Criminal Behavious and Mental Health, 29, 286-295.

Howden, S., Midgley, J. & Hargate, R. 2018. Violent offender treatment in a medium secure unit. Journal of Forensic Practice, 20, 102-111.

Jung, S., Daniels, M., Friesen, M. & Ledi, D. 2012. An examination of convergent constructs among level of service measures and other measures. Journal of Forensic Psychiatry and Psychology, 23, 601-619.

Jung, S., Ledi, D. & Daniels, M. K. 2013. Evaluating the concurrent validity of the HCR-20 scales. Journal of Risk Research, 16, 697-711.

Kashiwagi, H., Kikuchi, A., Koyama, M., Saito, D. & Hirabayashi, N. 2018. Strength-based assessment for future violence risk: A retrospective validation study of the structured assessment of protective factors for violence risk (SAPROF) Japanese version in forensic psychiatric inpatients. Annals of General Psychiatry, 17, 5.

Kennedy, H., O'Neill, C., Flynn, G. & Gill, P. 2010. Dangerousness understanding, recovery and urgency manual (the DUNDRUM quartet) V1.0.21. Dublin, Ireland: Central Mental Hospital, National Forensic Mental Health Service and Academic Department of Psychiatry, University of Dublin, Trinity College.

Kennedy, H., O'Neill, C., Flynn, G., Gill, P. & Davoren, M. 2013. Dangerousness understanding, recovery and urgency manual (the DUNDRUM quartet) V1.0.26. Dublin, Ireland: Central Mental Hospital, National Forensic Mental Health Service and Academic Department of Psychiatry, University of Dublin, Trinity College.

Lewis, K., Olver, M. & Wong, S. 2013. The Violence Risk Scale: Predictive validity and linking changes in risk with violent recidivism in a sample of high-risk offenders with psychopathic traits. Assessment, 20, 150-64.

Long, C., Fulton, B., Dolley, O. & Hollin, C. 2011. Social problem-solving interventions in medium secure settings for women. Medicine, Science and the Law, 51, 215-219.

Longdon, L., Edworthy, R., Resnick, J., Byrne, A., Clarke, M., Cheung, N. & Khalifa, N. 2017. Patient characteristics and outcome measurement in a low secure forensic hospital. Criminal Behaviour and Mental Health, 28, 255-269.

Magoub, N. 1988. "Bridging" therapy in hospital - and community-based psychiatric nursing care: A comparative study. PhD. thesis, Sheffield City Polytechnic, Sheffield.

Middleton, C., Mamuza, J., Martin, M. & Webster, C. 2002. Short Term Assessment of Risk (STAR); Abbreviated manual (version 1, consultation edition). Working paper in forensic mental health. Hamilton, Canada: Centre for Mountain Health Services, St. Joseph's Healthcare.

Mokkink, L., Prinsen, C., Patrick, D., Alonso, J., Bouter, L., de Vet, H. & Terwee, C. 2018. COSMIN methodology for systematic reviews of patient-reported outcome measures (PROMS). User Manual. Version 1. Amsterdam, The Netherlands: VU University Medical Centre.

Nicholls, T., Brink, J., Desmarais, S., Webster, C. & Martin, M. 2006. The Short-Term Assessment of Risk and Treatability (START): A prospective validation study in a forensic psychiatric sample. Assessment, 13, 313-327.

Nicholls, T., Petersen, K., Brink, J. & Webster, C. 2011. A clinical and risk profile of forensic psychiatric patients: Treatment team STARTs in a Canadian service. International Journal of Forensic Mental Health, 10, 187-199.

Nonstad, K., Nesset, M.., Kroppan, E., Pedersen, T., Nottestad, J., Almvik, R. & Palmstierna, T. 2010. Predictive validity and other psychometric properties of the Short-Term Assessment of Risk and Treatability (START) in a Norwegian high secure hospital. International Journal of Forensic Mental Health, 9, 294-299.

O'Dwyer, S., Davoren, M., Abidin, Z., Doyle, E., McDonnell, K. & Kennedy, H. 2011. The DUNDRUM quartet: Validation of structured professional judgement instruments DUNDRUM-3 assessment of programme completion and DUNDRUM-4 assessment of recovery in forensic mental health services. BMC Research Notes, 4, 229.

O'Shea, L. & Dickens, G. 2014. Short-Term Assessment of Risk and Treatability (START): Systematic review and meta-analysis. Psychological Assessment, 26, 990-1002.

Olver, M., Lewis, K. & Wong, S. 2013. Risk reduction treatment of high-risk psychopathic offenders: The relationship of psychopathy and treatment change to violent recidivism. Personality Disorders, 4, 160-167.

Phelan, M., Slade, M., Thornicroft, G., Dunn, G., Holloway, F., Wykes, T., Strathdee, G., Loftus, L., McCrone, P. & Hayward, P. 1995. The Camberwell Assessment of Need: The validity and reliability of an instrument to assess the needs of people with severe mental illness. British Journal of Psychiatry, 167, 589-595.

Prochaska, J. & DiClemente, C. 1982. Transtheoretical therapy: Toward a more integrative model of change. Psychotherapy: Theory, Research and Practice, 19, 276-288.

Reed, V. & Woods, P. 2002. The Behavioural Status Index [BEST-Index] - a 'life skills' assessment for selecting and monitoring therapy in mental health care. London, UK: Psychometric Press.

Richter, M., O'Reilly, K., O'Sullivan, D., O'Flynn, P., Corvin, A., Donohoe, G., Coyle, C., Davoren, M., Higgins, C., Byrne, O., Nutley, T., Nulty, A., Sharma, K., O'Connell, P. & Kennedy, H. 2018. Prospective observational cohort study of 'treatment as usual' over four years for patients with schizophrenia in a national forensic hospital. BMC Psychiatry, 18, 289.

Robinson, D., Reed, V. & Lange, A. 1996. Developing risk assessment scales in forensic psychiatric care. Psychiatric Care, 3, 146-152.

Ross, T., Reed, V., Fontao, M. & Pfaefflin, F. 2012. Assessing reliability, validity, and clinical utility of the best-index in measuring living skills among forensic inpatients.

International Journal of Offender Therapy and Comparative Criminology, 56, 385-400.

Ross, T., Woods, P., Reed, V., Sookoo, S., Dean, A., Kettles, A., Almvik, R., ter Horst, P., Brown, I., Collins, M., Walker, H. & Pfafflin, F. 2007. Selecting and monitoring living skills in forensic mental health care: Cross-border validation of the BEST-index. International Journal of Mental Health, 36, 3-16.

Ross, T., Woods, P., Reed, V., Sookoo, S., Dean, A., Kettles, A., Almvik, R., ter Horst, P., Brown, I., Collins, M., Walker, H. & Pfaefflin, F. 2008. Assessing living skills in forensic mental health care with the behavioural status index: A European network study. Psychotherapy Research, 18, 334-344.

Royal College of Psychiatrists. 2021. Health of the nation outcome scales (HoNOS). Royal College of Psychiatrists. Accessed 10 June 2021. Available at: https://www.rcpsych.ac.uk/events/in-house-training/health-of-nation-outcome-scales

Singh, J., Grann, M. & Fazel, S. 2011. A comparative study of violence risk assessment tools: A systematic review and metaregression analysis of 68 studies involving 25,980 participants. Clinical Psychology Review, 31, 499-513.

Thomas, S., Slade, M., McCrone, P., Harty, M., Parrott, J., Thornicroft, G. & Leese, M. 2008. The reliability and validity of the forensic Camberwell Assessment of Need (CANFOR): A needs assessment for forensic mental health service users. International Journal of Methods in Psychiatric Research, 17, 111-20.

Walker, H., Tulloch, L., Boa, K. & Ritchie, G. 2019. A multi-site survey of forensic nursing assessment. Journal of Forensic Practice, 21, 124-138.

Webster, C., Martin, M., Brink, J., Nicholls, T. & Middleton, C. 2004. Manual for the Short-Term Assessment of Risk and Treatability (START). Version 1.0 (consultation edition). Hamilton/Port Coquitlam, Canada: St. Joseph's Healthcare/Forensic Psychiatric Services Commission.

Webster, C., Martin, M., Brink, J., Nicholls, T. & Desmarais, S. 2009. Short-Term Assessment of Risk and Treatability (START). Version 1.1. Hamilton/Port Coquitlam, Canada: St. Joseph's Healthcare/Forensic Psychiatric Services Commission.

Webster, C., Nicholls, T., Martin, M., Desmarais, S. & Brink, J. 2006. Short-Term Assessment of Risk and Treatability (START): The case for a new structured professional judgment scheme. Behavioral Sciences and the Law, 24, 747-66.

Whittington, R., Bjorngaard, J., Brown, A., Nathan, R., Noblett, S. & Quinn, B. 2014. Dynamic relationship between multiple START assessments and violent incidents over time: A prospective cohort study. BMC Psychiatry, 14, 323.

Wilson, K., Freestone, M., Taylor, C., Blazey, F. & Hardman, F. 2014. Effectiveness of modified therapeutic community treatment within a medium-secure service for personality-disordered offenders. Journal of Forensic Psychiatry and Psychology, 25, 243-261.

Wing, J., Beevor, A., Curtis, R., Park, S., Hadden, J. & Burns, A. 1998. Health of the Nation Outcome Scales (HoNOS): Research and development. British Journal of Psychiatry, 172, 11-18.

Wong, S. & Gordon, A. 2000. Violence Risk Scale manual. Saskatchewan, Canada: Research Unit, Saskatchewan Regional Psychiatric Centre.

Wong, S. & Gordon, A. 2006. The validity and reliability of the violence risk scale: A treatment-friendly violence risk assessment tool. Psychology, Public Policy and Law, 12, 279-309.

Woods, P. & Reed, V. 1999. The Behavioural Status Index (BSI) some preliminary reliability studies. International Journal of Psychiatric Nursing Research, 5, 554-61.

Woods, P., Reed, V. & Collins, M. 2001a. Measuring communication and social skills in a high security forensic setting using the Behavioural Status Index. International Journal of Psychiatric Nursing Research, 7, 761-777.

Woods, P., Reed, V. & Collins, M. 2001b. Measuring risk in a high security forensic setting through the Behavioural Status Index. International Journal of Psychiatric Nursing Research, 7, 793-805.

Woods, P., Reed, V. & Collins, M. 2003a. Exploring core relationships between insight and communication and social skills in mentally disordered offenders. Journal of Psychiatric and Mental Health Nursing, 10, 518-525.

Woods, P., Reed, V. & Collins, M. 2003b. The relationship between risk and insight in a high-security forensic setting. Journal of Psychiatric & Mental Health Nursing, 10, 510-507.

Woods, P., Reed, V. & Collins, M. 2004. Relationships among risk, and communication and social skills in a high security forensic setting. Issues in Mental Health Nursing, 25, 769-782.

Woods, P., Reed, V. & Collins, M. 2005. The Behavioural Status Index: Testing a social risk assessment model in a high security forensic setting. Journal of Forensic Nursing, 1, 9-19.

Woods, P., Reed, V. & Robinson, D. 1999. The Behavioural Status Index: Therapeutic assessment of risk, insight, communication and social skills. Journal of Psychiatric Mental Health Nursing, 6, 79-90.

Yoon, D., Spehr, A. & Briken, P. 2011. Structured Assessment of Protective Factors: A German pilot study in sex offenders. Journal of Forensic Psychiatry and Psychology, 22, 834-844.